The Psychology of Construal in the Design of Field Experiments*

Elizabeth Levy Paluck & Eldar Shafir

Department of Psychology &
Woodrow Wilson School of Public and International Affairs
Princeton University

April 1, 2015

*Preliminary version, prepared for the NBER Conference on Economics of Field Experiments, organized by Esther Duflo & Abhijit Banerjee. Thank you to Robin Gomila, Sachin Banker, Peter Aronow, and Ruth Ditlmann for helpful comments. Address correspondence to epaluck@princeton.edu and eshafir@princeton.edu.

1 Introduction

Why have you picked up this book? A fair assumption is that you are reading because you care about good experimental design. To create strong experimental designs that test people's responses to some intervention, most researchers consider classically recognized behavioral motivations. It does not take an advanced degree in psychology to understand that several types of motivations could affect an individual's engagement with and honesty during your experimental paradigm. Such motivations include strategic self-presentation, suspicion, level of education or mastery, trust, and principles of utilitarian motives, least effort, and optimization. For example, minimizing the extent to which your results depend on high levels of participant suspicion, or decisions to do do the least amount possible, is important for increasing the generalizability and reliability of your results.

Psychologists agree that these motivations are important to consider when designing experiments. But they rank other behavioral drivers higher, drivers of individual behavior often ignored by other experimental researchers: consistency, identity, emotional states like pride, depression, and hunger, social norms, and perceptions of ideas like justice and fairness. Moreover, psychologists are keenly aware of features of the immediate situation that promote or diminish these behavioral drivers. The question for any experimentalist is: how do we figure out which behavioral drivers matter in any one particular experiment, and how they matter?

In this chapter we focus on the idea of construal, a foundational but under-appreciated principle that psychologists employ to understand behavior and to design experiments that can better approximate and isolate the causal dynamics that lead to the behavior of interest. Construal is defined as an individual's subjective interpretation of a stimulus, whether the stimulus is a choice set, a situation,

another person or group or people, a survey, or an experimental intervention (Ross and Nisbett, 1991). People's construal of a survey, then, is the personal and subjective meaning that they attach to the items as a set, which is also influenced by their presentation in the survey.

In the last two decades, psychological insights have been integrated into the study of economic perception and behavior, creating a literature called behavioral economics or applied behavioral science (Kahneman, 2013). One result is that many economists interested in behavior have a greater appreciation for the seemingly mundane immediate situational features that can promote or diminish behavioral drivers and thus the behaviors themselves. In short, most behavioral scientists from a range of disciplines can now tell you that the "situation matters." For example, removing small demands on a person's time, such as signing a form, can dramatically increase rates of behavior like signing up for a 401K plan or opting to be an organ donor (Thaler and Sunstein, 2008).

The interdisciplinary behavioral science literature has generated a great deal of advice on how to design an intervention, given that the situation matters –that individuals are sensitive to the timing, physical location, and framing of an intervention (Datta and Mullainathan, 2014; Shafir, 2013). And while psychology has been merged with economics to create a more "behavioral science", psychologists have pointed out that this field could never quite be "behavioralized" (Ross and Nisbett, 1991) meaning that we cannot take the mentalizing processes out of the Stimulus-Behavioral response pattern studied long ago by early behavioral psychologists who trained rats and pigeons to respond to lights and sounds (Skinner, 1960; Seligman, 1970). In other words, the stimuli that we create for our interventions and for our experimental tests of those interventions are never interpreted directly, as the experimentalist might intend.

There is, in this sense, a presumption in standard economic thinking that is really quite radical from the point of view of a psychologist. When a person in presented, say, with a choice between options A and B, what she chooses between are not A and B as they are in the world, but, rather, she chooses between A and B as they are represented by the 3-pound machine that she carries behind the eyes and between the ears. And that representation is not a complete and neutral summary, but rather a specific and constructed rendering – a construal.

Building on previous work that discusses how to design interventions based on an understanding of situational pressures and individuals' construal of those pressures (Datta and Mullainathan, 2014; Ross and Nisbett, 1991; Shafir, 2013), this chapter points out ways in which participants' construal of the stimuli in your experiment – everything from the behavior in question to the setting, the intervention, the deployment of the intervention, and the measurement tools – should affect the way you design and deploy your experiment.

Acknowledging subjective interpretation of the experiment is not the same as claiming to have no knowledge of participants' own construction of reality. Psychologists can provide many ways in which construal processes might be systematic and predictable. Nonetheless, one deep message is that experimentalists need to be modest about and to explicitly test assumptions concerning how participants view experimental interventions. Being aware of and taking steps to understand participants' construal in advance can help you to design and deploy the kind of field experiment that will shed light on the causal processes leading to the behavior in which you are interested.

You as the investigator, furthermore, are not excluded from the forces of construal. Toward the end of this chapter, we will also explore how your own construal of your experiment and of the

data can affect the way you interpret your results, conduct replications, and recommend elements of your intervention for scaling up or for institutional policies. We begin by providing an overview of construal: its definitions, functions, and some illustrative examples.

Principle of construal

At the turn of the 20th century and particularly during the two world wars, psychologists were moving away from Freudian focus on the personal histories and individual differences driving behavior and behavioral disorders. Kurt Lewin, a German psychologist and an èmigrè who eventually directed the Center for Group Dynamics at MIT, developed a situationally-driven alternative to Freud's claim that conflicting forces within the individual (the id, the ego, and superego), only available through the introspection of the individual and her therapist, could explain behavior and individual decision making. To understand and to facilitate the scientific study of behavior, Lewin proposed, we should look to the conflicting forces of the environment surrounding the individual (laws, family pressures, the physical environment), and how those forces push an individual and her self-proclaimed beliefs and desires into particular behavior choices.

Lewin called this type of analysis a social psychological analysis of the tension system – tensions between individual motivations and environmental pressures – and showed through a series of field experiments how leaders, workplace hierarchies and decisionmaking arrangements, peers' public behavior, and the physical proximity of particular resources could promote or inhibit a person's personal desires and beliefs, and change behavior in predictable ways. His early theorizing formed the foundation of modern social and cognitive psychology, and today guides the assump-

tions that psychologists make as they design and evaluate of behavior change and decision making experiments.

As Lewin was exploring the importance of situational pressures on behavior, some psychologists took this view to the extreme, including radical behaviorists like B.F. Skinner who felt that all behavior was a response to objective environmental forces learned over time. This view, while at first popular, proved to be inaccurate in the absence of one other principle of human thought that Lewin proposed as a critical part of the tension system analysis: construal. Environmental forces are not directly and objectively perceived by the individuals inhabiting each tension system, Lewin reasoned. Perception is a subjective process, which can happen in a considered, deliberate manner or in a fast and less conscious manner. Construal, the act of interpreting and attaching subjective meaning to forces like one's peers, leaders, group identities, choices, and the like, is also inherently variable – a stimulus may be interpreted by the same person one way at a certain time or in a certain situation, and differently in the next situation. Similarly, two people judging the same stimulus can construe it in different ways.

Some classic examples of how construal can affect judgments and behavior include the following:

- Judgments of a stimulus depend on how you construe the judgment relative to similar stimuli you have adapted to in your environment: A rule is perceived as strict when you are used to lax rules, and as lax when you are coming from a stricter rule environment. This is intuitive, and easily demonstrated through judging water temperature with your hand, just after you have plunged your hand first in a cold bucket or a hot bucket of water. Judgment will be

relative and not reflective of an absolute physical (or social) property. (CITE)

- Framing affects construal: Framing a monetary amount as a loss or a gain changes its construal, and the risk attitude it elicits (Tversky and Kahneman, 1981). More generally, any frame depicting a stimulus (an idea, choice, or behavior) as consistent with or as a departure from a perceived reference point shifts an individual's reception of the stimulus (Kahneman and Miller, 1986).

- Self-appraisal is made through social comparisons: Judgments about the self, including accomplishments, motivations, the strength of particular identities, and ideologies, are often made relative to other individuals present in the situation or other individuals mentioned in the question (Markus and Kunda, 1986; Morse and Gergen, 1970)

- Taxes and subsidies provoke unintended reactions, depending on individuals' construal of the behaviors they target: Individuals may interpret economic incentives as psychological taxes (i.e., demotivating) when the incentives subsidize behavior that is self-motivated; likewise, economic taxes may be interpreted as psychological subsidies (motivating) when they punish behavior that individuals have mixed feelings about or are trying to stop (Miller and Prentice, 2013).

- Peer pressure is effective not just because of conformity but because peers redefine the behavior in question: Individuals do not just adopt peer behavior, but also their peers' construal of the behavior or the situation. For example, when individuals observe peers ranking "politician" very positively vs. very negatively as a profession, the individual's own ranking of the term politician changes, not out of mimicry but because the individual has a different kind of politician in mind as a result of their peers' ranking (Asch, 1940).

- Global judgments color more specific ones and earlier information changes the meaning of later information: For example, global traits such as warmth can change the construal of a more specific trait like intelligence: the latter is interpreted as wisdom when a person is globally judged to be warm but as cunning when the person is thought of as cold. Also, learning about a teacher's argument with a student is interpreted differently if it is first vs. later revealed that the teacher was voted teacher of the year by his students (Ross and Nisbett, 1991).

- The source of a message colors the meaning of the message: Asch also showed in a classic study (Asch, 1948) that the quote "a little rebellion now and then is a good thing" was interpreted significantly differently by students for whom it was attributed to Thomas Jefferson vs. Lenin.

- Ideology changes which facts are noticed, believed, and remembered: Partisanship determines which facts individuals attend to, believe, remember, and understand when consuming news or other kinds of fact-based reports (Vallone et al., 1985).

- Construal affects how individuals assess the relative importance of various causal factors. While lay individuals (and researchers) reasonably search out three types of "data" to understand the causes of behavior in the world, including observations of distinctiveness (how specific is the behavior to this instance or individual), consistency (over time, is this behavior observed in this situation or for this individual), and consensus (how many other people behave this way or in this situation), individuals are often biased toward dispositional explanations for behavior that favor a person's character over situational explanations that favor the pressures of the environment (Kelley, 1973; Ross and Nisbett, 1991).

In the words of the cognitive psychologist Jerome Bruner, individuals who construe stimuli differently according to current levels of adaptation, frames, social comparisons, and present desires are "going beyond the information given" (Bruner, 1957). Psychologists see this subjective interpretation as a normal feature of human cognition, which can happen deliberately and consciously as well as spontaneously and unconsciously. That construal can be an automatic and unconscious process troubles our ability as investigators to ask directly about how an individual's interpretation might depend on their current circumstance. Indeeed, individuals do not usually have insight into the ways in which problem presentation, peers, and other Lewinian environmental pressures affect their construal.

Fortunately, psychologists have identified some "systematic factors [that contribute] to variability and instability of meaning" when individuals construe various stimuli (Ross and Nisbett, 1991, 69). Ross and Nisbett (1991) review the classic literature on various "tools of construal," which include knowledge structures like scripts, schemas, models, and heuristics that help individuals to quickly and with minimal effort make sense of other people, situations, choices, and other stimuli.

Schemas, for example, are mental representations containing knowledge about a group of related topics. Once a schema is activated, subsequent stimuli are interpreted using topics contained in that schema, with consequences for memory, decisionmaking, judgment, and behavior. A schema for "farm" can influence an individual's attention when considering a farm environment; she would spend more time paying attention to aspects of the farm that do not fit with her farm schema, like the appearance of an octopus. In this case, her schema will predict what piece of information about the farm she spends the most time considering, and also what she remembers about the farm (Loftus and Mackworth, 1978). Scripts, such as a script for how to behave at an academic conference, imply

even more specific knowledge structures about the order in which certain events should unfold and how an individual is expected to behave during each event, such as a discussion section, a coffee break, and an evening dinner with colleagues (Schank and Abelson, 2013).

Scripts, schemas, and heuristics (Gilovich et al., 2002) may be investigated as local tools of construal that exist within a certain population (such as people exposed to a certain kind of farm or academic conference) or as tools of construal for most people (such as the status quo bias against change, which seems to exist in many different populations (Eidelman and Crandall, 2009; Kahneman et al., 1991)). These various tools of construal improve individuals' speed of interpretation, and even if they guide behavior and judgments in directions that deviate from the predictions of rational actor models, they help to make resulting behaviors and judgments more predictable.

Laypeople and social science researchers often fail to appreciate the role of construal in guiding individuals' behavioral responses; instead, they tend to attribute choices that deviate from some rational prediction or norm to individuals' dispensational characteristics like personality, intelligence, or ideology. The literature on construal encourages the view that behavior is not necessarily a product of a person's character, but rather a window into how the person construes their choices or environment.

For experimentalists and for policy makers (the consumers of much of this research) all this should be of great importance. Behavior in experiments, and its interpretation, is determined not simply by the objective building blocks of the experiment, but by what participants know, want, attend to, perceive, understand, remember, and the like. Thus, experiments that are otherwise well designed, like well-intentioned interventions, can fail because of the way they are construed by the

participants, as well as by the investigators themselves. The difference between success and failure can sometimes boil down to a relatively benign and normatively immaterial change in presentation and subsequent construal, rather than a complex and costly rearrangement of experimental logic or procedure.

In the following pages we will show how psychologists understand construal as important to the design of an effective experiment. We offer a number of suggestions for how you as an investigator can attempt to understand your participants' construal of the stimuli in your experiment, or how you might reach what we term *shared construal* with your participants. The goal is to design and deploy a stimulus (intervention) in a field setting that participants will construe the way you intend them to.

By shared construal we do not mean that investigators and participants understand a behavioral problem or a choice set in the same way. Naturally, the experimenter will know things that the subject does not, and might arrange things in ways that escape the subject?s attention. What we mean by shared construal is that the investigator inhabits participants? perspective as best they can, as they are designing the experiment.

Psychologists think of designing experiments as a way of creating different counterfactual worlds for their participants to inhabit and respond to. As the saying goes, "I can explain it for you, but I can't understand it for you." The point, then, is to design a world that does not require the experimenter to explain things to the participant. The participant ought to understand the world afforded by an experimental condition in the way the experimenter intended. How to do this is no easy feat, and there is no foolproof recipe to follow. In the following sections we offer sug-

gestions for understanding participants' construal, as well as your own, as you conceptualize your intervention and experiment (piloting phase), as you design and deploy your intervention and measurement (design phase), and as you interpret your results and plan follow up experiments or scale ups (interpretation phase).

2 Pilot: seek shared construal of behavior and the situation between investigators and participants

Piloting often means testing out an experimental paradigm before the actual trial. But piloting can also be time set aside to understand a participant population's construal of the behavior in question and of the situations involved in your experimental paradigm prior to designing the full experiment. In this sense, piloting is a research and discovery stage about construal. It requires a high level of modesty on the part of the experimenter about what's driving people, before crafting an intervention to test the behavior in question.

Before designing the intervention or the experimental paradigm (i.e., the content of the manipulation or the set-up and deployment of the manipulation and measurement), it is important to first understand the underlying drivers of the behavior in question, in the particular setting of interest. What are the restraining forces that cause the behavior not to be enacted, or the compelling forces that drive the behavior at particular times or among particular people?

Redelmeier et al. (1995) were interested in why homeless adults in a southeast region of Toronto,

Canada, repeatedly visited the emergency room for care, up to 60 times per year, even when they were not given everything they needed. One common construal of this behavior from medical professionals and researchers was that the behavior was driven by homeless adults' neediness, and that if hospitals provided them with more care, this would only increase demand. The authors used survey data to understand the construal of homeless adults who attended the ER: nearly one third mentioned that they were treated rudely by hospital staff and nearly half reported that their needs were not met at the time of their visit. Crucially, 42% reported that they returned to the ER because of an unmet medical need.

Based on this revised construal, Redelmeier et al. (1995) hypothesized that it was also possible that increased care could address homeless adults' perceived satisfaction with their care and lessen the number of return visits. These possibilities informed their experimental design: a compassionate care condition run by volunteers, to provide randomly-assigned homeless adults with extra (though non-clinical) attention during their visit through friendly conversations and other kinds of rapport-building, and a baseline condition in which the other half of the selected sample were treated as per emergency room policy. In this case, the compassionate treatment, which directly addressed participants' construal of the situation, led to a 30% drop in repeated visits to the emergency room.

It is notable that the experiment excluded homeless adults that might be unresponsive to changes in treatment, including those who were acutely psychotic, unable to speak English, or were intoxicated or extremely ill. These insights, along with the insights regarding the participants' vs. the medical professionals' understanding of repeated visits to the hospital, were won through familiarity with the context of the experiment, a willingness to admit uncertainty in the original interpretation of the observed behaviors, along with some systematic data collection regarding the

participants' own construals.

Investigating participants' construal of the behavior in question ahead of the experiment may shift the intervention design, helping you to re-conceptualize what is at issue for your experiment. The emergency room experiment is one in which individuals with the "big picture" view on the situation, the hospital administration and medical professionals, had the wrong construal of the behavior. Piloting helped to uncover a different insight into the behavior, a point that is also made by the literature on intervention design (Datta and Mullainathan, 2014)). The lesson here is not that the intervention achieved the right construal or not, the simple observation that all along these highly experienced hospital professionals had the wrong construal.

Piloting to understand local construals of the behavior can also help to understand the way in which the control or comparison condition is crafted to create the most precise contrast that draws out the causal factor believed to be responsible for the behavior. In the Redelmeier et al Redelmeier et al. (1995) experiment, experimenters focused on the way clinical treatment was delivered – with compassion.

Piloting can also help you to understand more about potential participants' construal of the environments where you plan to convene your experimental manipulations or measurements. The choice of an intervention site ought not to be guided by logistical convenience alone (though this is often critical to the successful deployment of a field experiment). Psychological research on context effects suggests that the site of the experiment can often drive some aspect of the experimental results. This point is less often appreciated compared to others. Obviously, you spend lots of time designing your experimental intervention – say, a community meeting, or a letter. Once designed,

will you convene your community meeting in a church, or in an old school, or in a restaurant? Will you send your thoughtfully crafted letter to a person's home, or to their workplace address? At the beginning or the end of month?

By this point, it will not surprise you that psychologists believe these choices matter deeply for how your participants will construe your intervention and the issues addressed by your intervention. In the famous Milgram study, participants were ordered by an experimenter to apply (ultimately fake) electric shocks to another participant in the study when he failed at a memory task. In the version of the study run at Yale University, 65% of participants were fully obedient to the experimenter's commands in delivering the maximum level of shock; 48% of participants were fully obedient when the study was run at a nondescript office building in the nearby city of Bridgeport without a visible university affiliation (Milgram, 1974).

Consider also a study of context and behavior by Berger et al. (2008), who examined voting outcomes when voters were assigned to vote in churches vs. schools. First, using observational data, they estimate that voters were approximately .5 percentage points more likely to vote in favor of increasing education spending (by raising the state sales tax from 5.0% to 5.6%) when they had been assigned to vote in a school vs. a church. Second, using an experiment in which participants were initially shown images of either schools or office buildings before stating their policy preference, the authors suggest that the school context primed participants to think positively about education and to vote in its favor. This effect held even though none of the participants believed that that exposure to school images boosted their support for the increased sales tax to support education, "suggesting environmental stimuli can influence voting choice outside of awareness" (p. 8847).

3 Design: ensure the intervention design, measurement, and deployment achieve shared construal between investigators and participants

3.1 Intervention design and deployment

Do participants in your field experiment understand the content of your intervention in the same way that you do as the investigator? In a now classic study, Gneezy and Rustichini (2000) introduced fines for picking up children late from daycare in a random subset of a sample of daycare centers in Israel. A fine is normally understood as a deterrent to action, and we might predict that parents in the treatment daycares would be motivated to show up on time, given the increased economic costs of their delay.

Instead, it seems that parents perceived the fines to be what some psychologists have termed an "economic tax but a psychological subsidy" (Miller and Prentice, 2013). Parents in the daycares where fines were implemented were significantly *more* likely to pick up their children late, an effect that persisted for even after the fine was removed 17 weeks later. Gneezy and Rustichini (2000) and others have reasoned that the fine reshaped the parents' understanding of their environment. In particular, the contract between parents and daycare providers was unclear in the situation of pickups. The fine clarified the contract – picking up your child late now "costs" this amount of money. Parents willing to pay this price came late. Another way of stating these results is that parents initially construed on-time pickups as a moral obligation, to be a good parent and a good

daycare client. The fine was construed as a psychological subsidy, a release from this moral guilt. Parents released from this moral obligation now felt that they only had to pay, and did not need to feel guilty about a late pickup.

What about community members and other bystanders to your experimental intervention? One negative externality of a field experiment might be that other (non-targeted) people in your participants' social networks may construe the intervention in unintended ways, and influence the participants in your experiment. Ross and Nisbett (1991) describe the surprising results of the Cambridge Somerville study, in which at-risk boys were randomly assigned to receive or to not receive a bucket of treatments for an extended period of time during early adolescence, including after school and summer programming, tutoring, home visits, and more. In the forty-year follow up to the experiment, investigators found that treatment participants had no better outcomes than control participants, and in some aspects including adult arrests and mortality, treatment participants looked somewhat worse.

Ross and Nisbett (1991) reason that one potential explanation for this lack of observed response to treatment rests in the community's construal of and response to the intervention. For example, community members like coaches and ministers who might have naturally reached out to the at-risk boys may have perceived that the treated boys no longer needed the help of the community, and withdrew crucial support. Another possibility is that community members construed the treated boys as much worse "troublemakers" due to all of the increased outside attention that they received, and treated them as such. These are post hoc proposals, but plausible ones that remind us of the importance of understanding the community construal of an intervention, even when those community members are not directly implicated by the experimental manipulation.

Anticipating these different construals, and achieving shared construal of your intervention design and the way it is deployed[1] in the participant population and the surrounding community is no small task. The examples we used point to the necessity of running a small scale version of the intervention to invite reactions and construals of the intervention that are not merely hypothetical in nature. Or, as in the case of the daycare experiment, interview participants to see how they understand their current "contract" with the daycare with respect to coming late – what drives them to come late, and how do they think the daycare feels about late pickups.

Finally, although it arrives after the implementation of the intervention, all experiments should involve some form of a manipulation check, which assesses whether and what the participant understood and noticed about the intervention. Manipulation checks are used all of the time in psychological experiments, for descriptively understanding how participants perceived the intervention, but they are relatively rare outside of psychology. Manipulation checks can be much more than a simple determination of treatment delivery, for generating the estimated LATE given randomized intention to treat. They can give a picture of the participants' construal of the intervention, through questions like "what did the letter tell you?" or "who sent that letter, and why do you think they sent it?" after participants are sent letters about an opportunity for financial literacy training. More intrusive manipulation checks via surveys or interviews can happen for a small subsample of the target population, or during piloting.

[1] See also recent work by Haushofer and Shapiro (2013) on participant construal of the fairness of the process of random assignment.

3.2 Measurement of outcomes and processes

How do participants construe your measurement tools? Do they understand your survey questions in the way they were meant to be understood? Do community members assisting with archival data collection perceive the data collection to be appropriate, and do they share the investigator's belief that the records of interest represent accurate traces of the behaviors under study?

Although survey measures are considered second-best to unobtrusively measured behavioral outcomes, they are often desirable additional pieces of information or the only source of outcome measurement in institutionally weak or disorganized settings without good records of behavior. Fortunately, an enormous literature in psychology on psychometrics, heuristics, and biases provides a framework for understanding when participants' construal of survey questions can differ from that of the investigator's.

When participants read or listen to a series of questions, they do not answer each question in isolation from the others. Rather, research suggests that participants attempt to make global sense of the questionnaire, assessing its general purpose and its broad themes. For example, one of the most widely-used questionnaires in psychology is the Rosenberg Self-Esteem Scale, which features a series of survey items aimed at assessing an individual's self-esteem – none of which include the term *self-esteem*. Participants rate their agreement with items such as "On the whole, I am satisfied with myself," and "All in all, I am inclined to feel that I am a failure" (reverse scored). (This is just one of the ways researchers show that participants do not respond to individual questions in isolation, which we address more below.)

Robins et al. (2001) intuited that participants taking this scale would quickly construe the purpose of the scale to be the measurement self-esteem, and that a response to a direct question about self-esteem would be equally valid. They constructed an alternative questionnaire consisting of one item: "I have high self-esteem." Ratings of this single item correlated to the same degree as did the multi-item self-esteem questionnaire with a broad number of criterion measurements, including other self-evaluations and biases, mental and physical health, and peer ratings of the participant. The single-item survey also cut down on the number of complaints from participants about answering the same question multiple times, on the number of skipped questions or random responses, and other problems with the multi-item survey protocol.

To be fair, in many cases a more complex topic necessitates multiple items; our point here is that participants are not passive recipients of each question item, one by one. They actively interpret the questionnaire, attempting to understand its overall purpose and topic from the individual items. Their interpretations, of course may overlap to various degrees with the investigator's own understanding. Many psychologists use the technique of "cognitive interviewing" (Willis, 2004) to test participant's understanding of a questionnaire before broader deployment. This technique involves asking the participant to react aloud to each question, talking through their reaction to the question, their evaluation of the potential responses to the question, and why they are providing the answer they are providing.

Participants can also construe certain questions in meaningfully different ways, simply as a result of what comes to mind as a function, for example, of the ordering of questions. Schwarz and Xu (2011) inquired about drivers' enjoyment when driving luxury as opposed to economy cars. In one study, they asked University of Michigan faculty and staff which car they drove (brand, model,

and year) and subsequently, how they "usually" feel while commuting. Consistent with common intuition, the drivers reported more positive emotions the more valuable the car they drove. Thus, estimated mean scores for drivers' positive affect while commuting was significantly higher while driving cars corresponding to the Bluebook values of a BMW than of a Honda Accord.

A reversed order of questioning, however, paints a different picture. In this ordering, university faculty and staff were first asked to report how they felt during their most recent episode of driving to work, and only after they had reported their feelings, were they asked what car they drove. In this condition, the quality of the car driven, as indexed by (the natural log of) its Bluebook value, was thoroughly unrelated to the drivers' affective experience.

These and similar findings make a simple but important point: What is momentarily on people's mind can influence their construal. The car matters to reported judgments of enjoyment when it is on the driver's mind, but not otherwise. When asked to report how they usually feel while driving their car, drivers think about their car to arrive at an answer. But when the car goes unmentioned, it figures not at all.

In other cases, participants construe instructions differently than intended by the investigator, particularly when the instructions involve concepts that are only understood on a surface level by participants. Item substitution is a phenomenon that was observed in the classic Linda problem Tversky and Kahneman (1973)[2] gave participants a description of a fictitious graduate student shown along with a list of nine fields of graduate specialization. Here is a description:

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and

[2]see also Kahneman and Frederick (2002), for further discussion.

clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

One group of participants was given a representativeness (or similarity) question; others were given a probability question. Participants in a representativeness group ranked the nine fields of specialization by the degree to which Tom W. "resembles a typical graduate student" in each of those fields. Participants in the probability group ranked the nine fields according to the likelihood of Tom W. specializing in each. Figure 1 below plots the mean judgments of the two groups. The correlation between representativeness and probability is nearly perfect (.97), showing near-perfect attribute-substitution. Representativeness judgments – which are natural and automatic – are more accessible than probability judgments, which are not intuitive and can be rather difficult. (And there is no third attribute that could easily explain both judgments.) When asked about probability, people substitute similarity judgments for their response. This, of course, can lead to actual error, where things that are more similar, but less likely, are rated higher in likelihood. The study also showed that participants' own probability judgments correlated highly negatively with their own estimated base rates of the graduate fields of specialization.

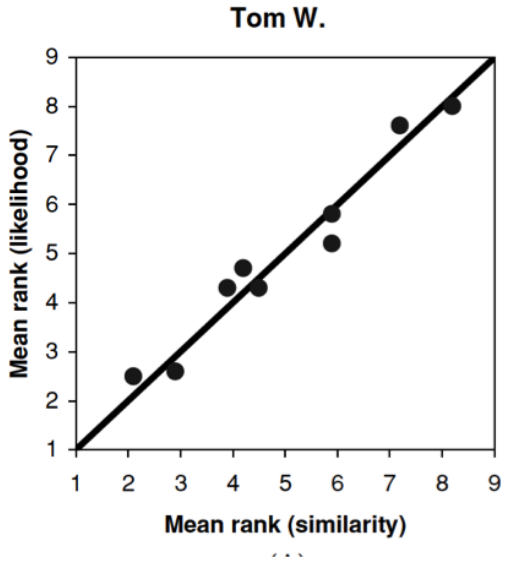

Figure 1: Tversky and Kahneman, 1973

Many investigators, particularly those working with less educated populations, use pictures to help with participant construal of the questionnaire – quite literally, pictures to illustrate the point. Participants use subjective interpretation with pictures as well as with words of course, and it is important to pilot how well the pictures communicate the point of the question or of the response option. One of us used a pictorial scale of depression for a field experiment conducted in Rwanda. The scale had been previously used in published work in the same country and more broadly in the Great Lakes region of Africa. It asked participants to answer the question "how have you been feeling in the past few weeks" by pointing to one of a series of pictures featuring a person carrying a stone. From picture to picture, the stone increased in size: on one end of the scale, the person held a small stone in his hand, and at the other end of the scale, the person was bent in half as they held up the weight of an enormous boulder on their shoulders.

Because the scale had been used successfully in previous studies in the area, we brought the scale directly into the field without a pilot. At one field site, the first participant was asked how they were currently feeling, and was shown the pictorial scale. The participant waited, and then left the interview to confer with others nearby. When he returned, he informed the enumerator that he was willing to carry some of the smaller stones for him, but not some of the larger ones. The misunderstanding of the scale ran even deeper than this. The scale caused active discussions in this community, and we were informed that during the recent civil conflict a military group asked a group of young men from the community to help carry supplies for them, and the young men were never seen again. A scale to measure depressive reactions to trauma was construed by the community to relate to one of their original sources of trauma. We took care to clarify our intentions and to repair the situation with the community, threw out the scale from our study, and resolved to never again to use a scale without a pilot.

3.3 Investigator presence

How do participants construe who you are, as an investigator, and what your presence in their community means for them and for their participation in the experiment? Some ethical discussions encourage investigators to stay away from data collections or intervention deployments because participants' respect for or fear of scientists may lead them to construe their participation or their responses to certain types of questions as mandatory (Orne, 1962; Rosnow and Rosenthal, 1997).

Paluck (2009) reports that varying levels of government scrutiny and physical security in the post-conflict countries where she has deployed field experiments has led to different self-presentation

strategies for enumerators and other representatives of the experiment. For example, in Rwanda, where security was excellent and government scrutiny was extremely high, enumerators identified themselves strongly with the university supporting the investigator and the study. However, just across the border in the Democratic Republic of Congo (DRC) where security and government surveillance were low, enumerators wore t-shirts featuring the local NGO that was collaborating with the university. In Rwanda, participants would have construed the emphasis on the NGO to mean that their responses were subject to government surveillance, as were most NGOs in the country during the experiment. However, in DRC, participants needed reassurance of the legitimacy of the experiment from a known local source, the NGO, due to the lack of security, and did not construe the NGO as an actor that would share their answers with the government.

Many other examples are possible, but here our bottom line is that the perceived source of the experiment will affect participant's construal of their choice in participating or not, the confidentiality of their responses, and the overall meaning of the experiment, among other things. We may even use the analogy of your own construal of the source of information in this chapter: as an economist reading this chapter, might you find it more authoritative if you knew it was coming from two economists, rather than two psychologists?

4 Interpret: how do investigators construe what matters in the data?

Thus far, our focus has been on how participants construe the various stimuli in your experiment. But investigators use the same tools of construal as participants, which means that we construe what participants tell us in ways that may not match up with their actual meanings. In other words, we may find ourselves in the role of Gary Larsen's dog Ginger, hearing only the information that we deem relevant, as participants tell us about themselves (see Figure 2).

Construing participants' self reports is not the only way that construal processes operate for investigators, shaping the way in which we understand the outcomes of our experiments. Construal can affect the way in which we conduct data analysis, the factors we interpret to be important for an experimental replication, and also for scaling up the intervention.

Recently, social scientists have laid out a rationale and evidence for the advantages of pre-registration of analyses prior to the deployment of a field experiment or to the commencement of analysis (Casey et al., 2011; Olken, 2015; Committee, 2015). Just as (Vallone et al., 1985) pointed out that partisanship can affect what individuals see in a factual news article, so too can researchers selectively pick analyses that support their preferred hypothesis in a large dataset (Casey et al., 2011). As Olken puts it, "Even researchers who have the noblest of intentions may end up succumbing to the same sorts of biases when trying to figure out how, ex-post, to make sense of a complex set of results" (p. 1).

Psychologists understand this practice as a result of the ordinary and sometimes inevitable pro-

Figure 2: Investigators listening to participants: Are we hearing the meaningful parts of their message? Copyright, The Far Side.

cess of construal: what you understand to be the most important test at the design stage can change as you observe the process of data collection, as you analyze your data, and as you form a working hypothesis about the study results. While there are non-negligible costs to pre-registering all of your analyses in advance (Olken, 2015) there are also clear advantages. In addition to publicly committing to a priori predictions, preregistration can help investigators think more carefully about their hypotheses as they design and modify the experimental protocol. Another practice that can help the post hoc downweighting of experimental hypotheses is registering a field experiment. This practice helps to prevent the selective reporting of entire trials that do not yield the results expected by investigators (using, for example, `http://www.socialscienceregistry.org/` or the newly-instantiated Open Science Framework).

Construal also potentially shapes which factors investigators take to be the generalizable lesson of the overall experiment: i.e., what was the causal driver of the results? At first blush, this may seem counterintuitive. Randomization of an independent variable allows for the estimation of a causal relationship. But how do investigators interpret what exactly was the important feature of the independent variable, in order to replicate the causal relationship using experiments in different contexts, or to scale up their study?

Consider the field experiment that Bertrand et al. (2010) ran in South Africa, in which they manipulated information a bank provided about loans in letters to their clients. Some of the information provided was central to what clients should want to know about the loans, including interest rates and the number of different types of loans. Other information was more peripheral, such as a picture of a man vs. a woman that was embedded in the letter's graphic design. The researchers found that having a picture of a woman on the letter significantly increased demand for the loan, an

effect size that was worth 25% of a reduction in the loan interest rate.

What is the conclusion of this experiment? Women's pictures increase loan take-up? Should we always expect pictures of women to increase the take-up of financial products? Would pictures of women work equally well in Belgium, or would other kinds of pictures be more attractive in that country? How these investigators construe the meaning of the woman's picture as the causal driver of loan take-up determines how they might try to replicate the experiment in other contexts, or how they might want to institutionalize or scale-up their results for the specific bank they worked with in South Africa. Replicating experiments with other institutions, whether they are different banks or governments or other firms, also creates the possibility that participants will construe the intervention very differently when it originates from a very different source. It is the investigator's job to attempt to distill what was most important about the original successful experiment for the replication or scale-up.

Replicating experiments

At its most general level, the South Africa loan experiment teaches us that simple, seemingly peripheral tweaks to advertisements of financial products can make a big difference. Apart from this, it may be unclear how to construe the specifics of the manipulation, regarding the photo of the attractive woman. Our general advice is to think about the conversion of specific manipulations in an experiment like you would the conversion of currency – i.e., if you conduct a behavioral experiment using Shekels in Israel, you would not think twice about converting to using Yen in Japan. This advice is obvious, but many field experimental replications are often encouraged to replicate

surface structure without replicating deep structure.

What do we mean by surface vs. deep structure? Take a garden path sentence and one that clarifies its meaning:

The horse raced past the barn fell

The horse that was raced past the barn fell

The first sentence, a classic garden path sentence, is grammatically correct but leads the reader to parse the sentence in a way that ultimately renders the sentence meaningless, by the time she is finished reading (the meaning is indicated by the second sentence, with additional words that allow for a correct interpretation of the phrase). While many replications may be "grammatically correct," they do not communicate the same meaning to their participants. Their paradigms have not been converted into the local currency. These analogies will hopefully give pause to any investigator who is interested in a "direct replication" of a study. A replication needs to replicate the deep structure, not the surface structure, i.e. it should replicate the participants' construal of the original study. This may take a more radical reconfiguring of the experimental stimuli, since construal of those might differ from one context to the next.

And speaking of designing the experimental stimuli for a replication study, we would be remiss if we did not offer a few more phrases by way of advice.

Let's eat, Grandma!

Let's eat Grandma!

That is, small nuances can save lives. Or change the way that your experimental stimuli are

construed during a replication.

All that we have discussed so far should convince you that experiments are not off-the-shelf tests. While we have no surefire method for replication, our advice is to replicate the participants' construal, not just the stimuli used in the original study. Repeat the psychologically important structure, not the superficial (and potentially garden path) structure of an experiment. For recent discussions on conceptual replication from psychology, see Monin et al. (2014).

Institutionalizing and scaling up experimental results

Discerning how to construe the causal drivers of your effect for a replication presents similar challenges to those encountered when attempting to identify which factors should be "scaled-up." By scale-up, we mean either a large-scale replication of your experiment or the installation of your experimental manipulation as part of a public or private institution's regular operating procedure. Moreover, scaling up your experimental manipulation introduces an additional complication, which is that your targeted population will most likely receive the experimental stimulus from a source that is different from the source in the original experimental evaluation. As we have already discussed, the source (e.g., university, non-governmental organization or government agents) matters a great deal for the participant's construal of the intervention, and may change the efficacy or the strength of the intervention effect. Additionally, interventions that are no longer presented as a trial, or as "experimental" may be construed differently when they are presented as a policy or a standard operating procedure.

To our knowledge, one of the most striking and sobering examples of a shift in participant's

construal from an experimental to an institutionalized policy is the domestic violence experiment led by the National Institute of Justice (Garner et al., 1995). The experiment used an encouragement design for police officers responding to a call reporting a domestic incident. Officers were assigned to arrest, mediate, or separate upon arrival at the scene through a colored notepad, though they could break with the randomization in the case of an emergency. The estimated effect of this experiment revealed the importance of arrests for preventing recidivism in domestic abuse, which reduced estimated future violence by more than 50%. The results were subsequently used to justify laws promoting arrests of individuals believed to be responsible for spousal abuse. Subsequently, Iyengar (2010) provided estimates showing that these laws have *increased* the number of intimate partner homicides where they have been implemented.

Setting aside debates about the methods and findings from Iyengar (2010) vs. those from the National Institute of Justice experiments, we ask how laws mandating arrest of abusive spouses could change violence survivors' construal of a call to the police for help. During the National Institute of Justice's experiment, a call to the police was understood as a call for help, without a clear idea of what the police officer might do once on the scene. Clearly, abusive partners, would never construe a call to the police as a welcome action; however prior to laws favoring immediate arrest, these calls were not understood as a partner requesting an arrest. Once inscribed into law, a call to the police meant a call to arrest the partner. Both partners in a domestic dispute presumably shared this new construal, which could explain why homicides rose following the laws.

In sum, a target population's understanding of an intervention may change as the intervention scales, comes from a different source, slightly changes form, and is no longer novel. Thinking about participants' construal in this way is a means of brainstorming the negative externalities of a scale-

up. Participant construal provides a rubric for understanding and anticipating negative externalities.

5 Concluding thoughts

*Forthcoming following discussion at conference.

Questions for our discussant: are there ways of communicating these ideas to economists that make clearer the connections between economic approaches to experiments and design problems, and the way that construal can address these problems?

References

Asch, Solomon E (1940), "Studies in the principles of judgments and attitudes: Ii. determination of judgments by group and by ego standards." *The Journal of Social Psychology*, 12, 433–465.

Asch, Solomon E (1948), "The doctrine of suggestion, prestige and imitation in social psychology." *Psychological review*, 55, 250.

Berger, Jonah, Marc Meredith, and S Christian Wheeler (2008), "Contextual priming: Where people vote affects how they vote." *Proceedings of the National Academy of Sciences*, 105, 8846–8849.

Bertrand, Marianne, Dean S Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman (2010), "What's advertising content worth? evidence from a consumer credit marketing field experiment." *Quarterly Journal of Economics*, 125, 263–306.

Bruner, Jerome S (1957), "Going beyond the information given." *Contemporary approaches to cognition*, 1, 119–160.

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2011), "Reshaping institutions: Evidence on aid impacts using a pre-analysis plan." Technical report, National Bureau of Economic Research.

Committee, The TOP Guidelines (2015), "Promoting an open research culture: The top guidelines for journals." *Working paper*, 1, 1–2.

Datta, Saugato and Sendhil Mullainathan (2014), "Behavioral design: A new approach to development policy." *Review of Income and Wealth*, 60, 7–35.

Eidelman, Scott and Christian S Crandall (2009), "A psychological advantage for the status quo." *Social and psychological bases of ideology and system justification*, 85–106.

Garner, Joel, Jeffrey Fagan, and Christopher Maxwell (1995), "Published findings from the spouse assault replication program: A critical review." *Journal of Quantitative Criminology*, 11, 3–28.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (2002), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Gneezy, Uri and Aldo Rustichini (2000), "A fine is a price." *Journal of Legal Studies*, 29, 1–18.

Haushofer, Johannes and Jeremy Shapiro (2013), "The social costs of randomization."

Iyengar, Radha (2010), "Does arrest deter violence? comparing experimental and nonexperimental evidence on mandatory arrest laws." In *The Economics of Crime: Lessons For and From Latin America* (Rafael Di Tella, Sebastian Edwards, and Ernesto Schargrodsky, eds.), 421–452, NBER/University of Chicago Press.

Kahneman, Daniel (2013), "Foreword." In *The behavioral foundations of public policy* (Eldar Shafir, ed.), 7–9, Princeton University Press.

Kahneman, Daniel and Shane Frederick (2002), "Representativeness revisited: Attribute substitution in intuitive judgment." *Heuristics and biases: The psychology of intuitive judgment*, 49.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler (1991), "Anomalies: The endowment effect, loss aversion, and status quo bias." *The journal of economic perspectives*, 193–206.

Kahneman, Daniel and Dale T Miller (1986), "Norm theory: Comparing reality to its alternatives." *Psychological review*, 93, 136–153.

Kelley, Harold H (1973), "The processes of causal attribution." *American psychologist*, 28, 107.

Loftus, Geoffrey R and Norman H Mackworth (1978), "Cognitive determinants of fixation location during picture viewing." *Journal of Experimental Psychology: Human perception and performance*, 4, 565.

Markus, Hazel and Ziva Kunda (1986), "Stability and malleability of the self-concept." *Journal of personality and social psychology*, 51, 858.

Milgram, Stanley (1974), *Obedience to authority*.

Miller, Dale T and Deborah A Prentice (2013), "Psychological levers of behavior change." In *The behavioral foundations of public policy* (Eldar Shafir, ed.), 301–309, Princeton University Press.

Monin, Benoît, Daniel M Oppenheimer, Melissa J Ferguson, Travis J Carter, Ran R Hassin, Richard J Crisp, Eleanor Miles, Shenel Husnu, Norbert Schwarz, Fritz Strack, et al. (2014), "Commentaries and rejoinder on klein et al.(2014)."

Morse, Stan and Kenneth J Gergen (1970), "Social comparison, self-consistency, and the concept of self." *Journal of personality and social psychology*, 16, 148.

Olken, Ben (2015), "Pre-analysis plans in economics." *Journal of Economic Perspectives*.

Orne, Martin T (1962), "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American psychologist*, 17, 776.

Paluck, Elizabeth Levy (2009), "Methods and ethics with research teams and ngos: Comparing experiences across the border of rwanda and democratic republic of congo." *Surviving Field Research: Working in Violent and Difficult Situations*, 38–56.

Redelmeier, Donald A, Jean-Pierre Molin, and Robert J Tibshirani (1995), "A randomised trial of compassionate care for the homeless in an emergency department." *The Lancet*, 345, 1131–1134.

Robins, Richard W, Holly M Hendin, and Kali H Trzesniewski (2001), "Measuring global self-esteem: Construct validation of a single-item measure and the rosenberg self-esteem scale." *Personality and social psychology bulletin*, 27, 151–161.

Rosnow, Ralph L and Robert Rosenthal (1997), *People studying people: Artifacts and ethics in behavioral research*. WH Freeman New York.

Ross, Lee and Richard E Nisbett (1991), *The person and the situation: Perspectives of social psychology.* Mcgraw-Hill Book Company.

Schank, Roger C and Robert P Abelson (2013), *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schwarz, Norbert and Jing Xu (2011), "Why don't we learn from poor choices? the consistency of expectation, choice, and memory clouds the lessons of experience." *Journal of Consumer Psychology*, 21, 142–145.

Seligman, Martin E (1970), "On the generality of the laws of learning." *Psychological review*, 77, 406.

Shafir, Eldar (2013), *The behavioral foundations of public policy*. Princeton University Press.

Skinner, Burrhus F (1960), "Pigeons in a pelican." *American Psychologist*, 15, 28.

Thaler, Richard H and Cass R Sunstein (2008), *Nudge*. Yale University Press.

Tversky, Amos and Daniel Kahneman (1973), "Availability: A heuristic for judging frequency and probability." *Cognitive psychology*, 5, 207–232.

Tversky, Amos and Daniel Kahneman (1981), "The framing of decisions and the psychology of choice." *Science*, 211, 453–458.

Vallone, Robert P, Lee Ross, and Mark R Lepper (1985), "The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre." *Journal of personality and social psychology*, 49, 577.

Willis, Gordon B (2004), *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.